연관 태그의 군집화를 위한 클러스터링 기법 비교 연구*

A Comparative Study on Clustering Methods for Grouping Related Tags

한 승 희(Han, Seunghee)**

목 차

초 록

본 연구에서는 태그 공간에서 정보의 효율적 탐색을 위해 이용자에게 제공될 수 있는 연관 태그 클러스터의 생성을 위해 다양한 유사계수와 클러스터링 기법을 적용한 후 그 결과를 평가하고 비교 분석함으로써 연관 태그의 클러스터링에 가장 적합한 클러스터링 알고리즘을 확인하고자 하였다. Delicious에서 임의의 태그 10개를 대상으로 각각 300개의 문서에서 추출한 연관 태그를 대상으로 태그쌍 간의 연관성을 측정한 후 계층적 기법과 비계층적 기법을 적용하여 생성된 클러스터를 대상으로 클러스터 적합도를 측정한 결과, 일반적으로 용어 클러스터링에서 널리 활용되는 것으로 알려진 워드 기법이 코사인 유사계수와 결합했을 때 거의 모든 실험 대상에 대해 유사한 경향을 보이면서 가장 우수한 성능을 나타내는 것으로 나타났다. 연관 태그 클러스터는 정보관리 측면에서 유사한 합목적성을 갖는 태그끼리 군집을 이루면서 용어의 중의성을 해소함으로써 태그 공간에서의 이용자의 정보 탐색에 유용하게 활용될 것이다.

ABSTRACT

In this study, clustering methods with related tags were discussed for improving search and exploration in the tag space. The experiments were performed on 10 Delicious tags and the strongly-related tags extracted by each 300 documents, and hierarchical and non-hierarchical clustering methods were carried out based on the tag co-occurrences. To evaluate the experimental results, cluster relevance was measured. Results showed that Ward's method with cosine coefficient, which shows good performance to term clustering, was best performed with consistent clustering tendency. Furthermore, it was analyzed that cluster membership among related tags is based on users' tagging purposes or interest and can disambiguate word sense. Therefore, tag clusters would be helpful for improving search and exploration in the tag space.

키워드: 태그 클러스터링, 용어 클러스터링, 연관 태그, 태그 기반 정보검색
Tag Clustering, Term Clustering, Related Tags, Tag-based Retrieval

---

* 이 논문의 일부는 2009년도 제16회 한국정보관리학회 학술대회에서 발표된 논문을 수정·보완한 것임.
이 논문은 2009학년도 서울여자대학교 교내학술특별연구비의 지원을 받았음.

** 서울여자대학교 사회과학대학 문헌정보학과 조교수(hanshee@swu.ac.kr)
논문접수일자: 2009년 8월 24일 최초심사일자: 2009년 8월 24일 게재확정일자: 2009년 9월 7일
한국문헌정보학회지, 43(3): 399-416, 2009. [DOI:10.4275/KSLIS.2009.43.3.399]

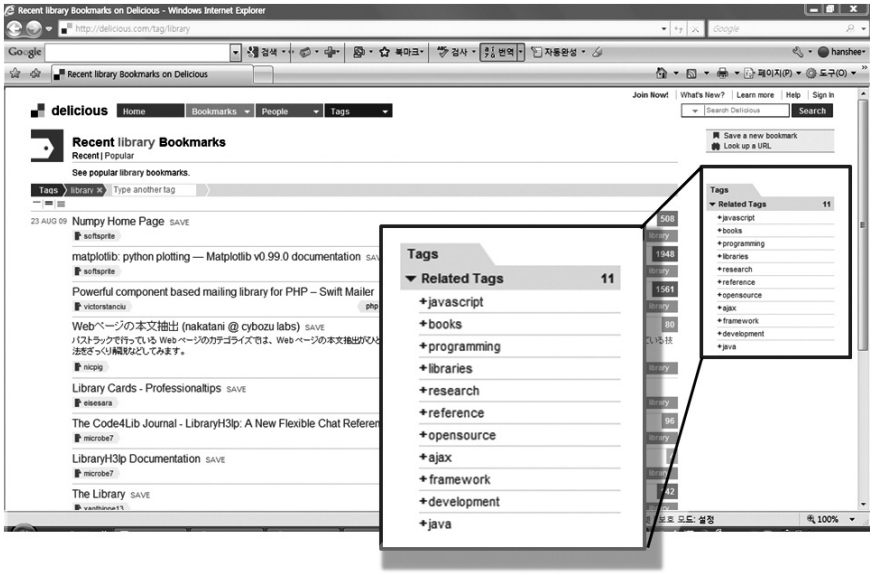
1. 서 론

최근 인터넷 환경이 웹 2.0 환경으로 진화하면서 이용자의 참여와 협력이 웹의 진화와 생성에 가장 중요한 역할을 하고 있는데, 이러한 이용자 참여형 웹 2.0의 형태 중 가장 대표적인 것이 바로 태그 메커니즘에 의한 협력적 태깅(collaborative tagging)이라고 할 수 있다. Flickr(http://flickr.com)나 Delicious (http://delicious.com)와 같이 태그를 활용한 서비스의 등장과 성공은 이용자 참여를 기반으로 한 데이터 진화의 좋은 예라고 할 수 있다.

협력적 태깅 서비스의 확산은 이용자의 정보관리 측면에도 큰 변화를 가져왔다. 우선, 정보조직의 측면에서 살펴보면, 미리 정해놓은 분류체계에 기반한 택사노미 중심에서 벗어나 이용자가 자유롭게 부여한 키워드에 기반한 폭소노미 형태로 진화함으로써 유연하고 역동적으로 정보를 분류할 수 있는 구조를 제공하게 되었다. 이러한 폭소노미 기반의 분류체계로의 변화는 정보의 검색에도 변화를 가져왔다. 전통적인 정보검색 환경에서의 통제어휘와 달리 태그는 이용자로 하여금 정보에 대한 직관적인 이해를 기반으로 정보의 검색이 가능하도록 돕는다(Hammond et al. 2005). 뿐만 아니라 이용자들은 브라우징 과정에서 의도하지 않은 우연한 발견(serendipity)을 통해 정보를 획득할 수 있게 되었다(Mathes 2004).

그러나, 태그를 이용한 정보관리 환경의 변화가 정보검색 영역의 근본적인 문제점을 해결한 것은 아니다. 실제로 대다수의 태깅 서비스에서는 태그를 이용한 검색 서비스를 제공하고 있으나, 이용자가 입력한 질의어와 태그를 매칭시키는 키워드 검색 방식으로 검색 결과를 제공하고 있어, 현재로서는 태그 데이터가 본질적으로 가지고 있는 유연성과 역동성을 활용하여 정보를 검색하는 것이 거의 불가능하다. 이러한 문제점은 태그가 가지고 있는 비통제어휘적 특성에서 기인한다고 할 수 있다(Fichter 2006; 이재윤, 정도헌 2008). 즉, 여러 이용자가 다양한 방식으로 자유롭게 태그를 부여하기 때문에 태그 데이터에는 모호성과 비일관성이 내재되어 있고, 태그 간의 관계를 수평적인 형태로만 표현할 수 있기 때문에 태그 간에 내재된 의미관계를 나타내기 어렵다는 단점을 가지고 있다. 그리고 이러한 단점은 정보검색에서 해결해야 할 과제 중 하나인 의미모호성, 다의어, 동의어, 이형 등과 같은 언어의 중의성 해소와 관련된 것으로, 비통제어휘가 갖고 있는 문제와 거의 동일하다.

태그 검색이 갖고 있는 문제점을 해결하기 위해 일부의 서비스에서 연관 태그(related tags)를 이용자에게 제공함으로써 태그 공간에서의 탐색의 효율성을 높이고자 하였다. 예를 들어, Delicious에서는 〈그림 1〉과 같이 특정 태그와 연관이 있는 태그들을 이용자에게 제공함으로써 이용자가 연관된 개념에 대해서도 정보를 탐색할 수 있도록 돕고 있다. 연관 태그는 이용자로 하여금 탐색하고자 하는 주제에 대해 연관 개념으로의 인지적 확장을 가능하게 하므로 탐색의 유연성을 제공하는 역할을 수행하고 있다. 이러한 관점에서 최근 일부 연구자들은 태그 공간의 탐색 효율성을 향상시키기 위해 이용자에게 태그 데이터를 효과적으로 제공하는 방법에 대해 관심을 갖기 시작하였다(Begelman, Keller, and Smadja 2006; 박병재, 우종우 2008;



〈그림 1〉 Delicious의 연관 태그: 'library' 사례

이순규, 김정훈, 이지형 2008; 이시화, 이만형, 황대훈 2008; Candan, Caro, Sapino 2008; Shepitsen et al. 2008; Simpson 2008; Schrammel, Leitner, and Tscheligi 2009).

이 연구에서는, 태그 공간에서의 효율적인 정보 탐색을 위해 연관 태그를 이용자에게 효과적으로 제공하기 위한 방법 중 클러스터링 기법을 제안하고, 연관 태그의 클러스터 생성에 가장 적합한 기법을 확인하기 위해 다양한 클러스터링 기법을 적용한 후, 그 결과를 평가하여 비교・분석하고자 한다.

2. 태그 클러스터링

2.1 태그 기반 정보검색의 한계와 해결책

태깅 서비스에서 이용자에게 연관 태그를 제공하는 것은 새로운 정보를 탐색할 수 있는 기능뿐만 아니라 연상되는 단어를 통해 기억나지 않은 검색 대상을 찾아주는 기능을 한다. 즉, 검색하려고 하는 검색 대상의 명확한 정보를 모르는 상태에서 기억나는 단편적인 정보를 통해 검색 대상을 찾을 수 있다. 앞 절에서 언급한 바와 같이 태그는 매우 유연하고 역동적인 분류체계를 제공한다. 그러나 유연성과 역동성의 확보로 인해 발생하는 근본적인 한계를 가지고 있는 것이 사실이다. 현재 소셜 태깅 서비스에서 제공하고 있는 태그 기반 정보검색의 한계점은 다음과 같다(Begelman, Kellier, and Smadja 2006).

2.1.1 탐색(search)의 제한

일반적으로 사람들은 자신의 배경지식이나 인지구조적 특성에 의해 하나의 사물에 대해 연상하는 내용이 다양하다. 이러한 관점에서

볼 때, 다수 이용자의 협력을 기반으로 특정 정보에 대한 키워드의 연상으로 이루어지는 소셜 태깅의 결과는 다양하게 나타날 수밖에 없다. 예를 들어, '자바'라는 단어가 주어졌을 때, 어떤 이용자는 '프로그래밍'을 떠올릴 수 있지만, 또 다른 이용자는 '커피'를 떠올릴 수 있다. 이러한 태그 공간의 다양성은 오히려 태그 공간의 효율적 탐색에 방해가 되는 잡음이 되어, 다른 이용자에 의해 태깅된 정보를 탐색하는 것을 어렵게 만드는 요소로 작용한다.

2.1.2 구독(subscription)의 제한

웹 2.0 환경에서 제공하고 있는 서비스 중 하나인 RSS 서비스는 이용자의 관심 주제에 따라 새로운 정보를 손쉽게 얻을 수 있기 때문에 많은 이용자들이 활용하고 있다. 일반적으로 RSS 서비스 이용자들은 특정 주제에 대한 정확한 자료(정확률)보다는 많은 자료(재현율)를 얻고자 하는 경향이 있다. 그러나 현재의 소셜 태깅 서비스를 기반으로 한 RSS 서비스의 경우에는 재현율을 향상시키는 것이 어렵다. 예를 들어 한 이용자가 '자바'와 '기사'라는 두 태그를 이용하여 RSS 구독을 신청했다면, 이 이용자는 '기사'와 유사 개념이라 할 수 있는 '블로그' 등으로 태깅된 정보에 대해서는 구독할 수 없다.

2.1.3 탐험(exploration)의 제한

일반적으로 태그 공간을 탐험하는 방법에는 두 가지가 있다. 하나는 태그를 키워드로 이용하여 검색하는 것이고, 다른 하나는 태그 클라우드와 같이 시각화된 태그 공간을 탐색하는 것이다. 그러나 이 두 방법 모두 이용자의 정보 요구를 만족시키기에는 문제점이 있다. 우선, 태그 검색의 경우 일반 검색엔진의 키워드 검색결과와 다른 점을 찾기 어렵다. 또한 태그 공간을 탐색하는 경우에 있어서는 태그 클라우드에서는 두 개 이상의 태그를 동시에 활용하여 검색의 범위를 좁히거나 넓히는 것이 불가능하며, 카테고리 기반 검색엔진과 같이 태그 정보에 대해 계층적으로 접근할 수 없어 정보검색이 효율적으로 이루어지지 못하고 있다.

현재의 태깅 서비스에서는 이용자마다 다양하게 부여되는 태그 데이터가 수평적인 관계로 표현되며, 탐색 시에도 두 개 이상의 연관 태그를 조합함으로써 개념을 확장하는 것이 거의 불가능하다. 이러한 제한점은 현재의 태깅 시스템이 기본적으로 단어 간의 어휘적 관계(lexical relation)를 허용하지 않고 있기 때문인 것으로 해석할 수 있다. 일반적으로 단어는 서로 독립적이지 않으며 연관되어 있기 때문에 이러한 단어 간의 어휘적 관계를 이해하는 것이 올바른 탐색을 위한 중요한 요소가 된다. 일부의 태깅 서비스에서 특정 태그에 대한 연관 태그를 제시함으로써 어휘적 관계를 이용한 탐색의 확장을 지원하고 있으나, 기본적으로 태그 데이터는 비구조화된 데이터 형태를 가지고 있기 때문에 연관 태그만으로는 다양한 어휘적 관계를 표현하기 어렵다.

비구조화된 태그 데이터를 구조화함으로써 앞에서 언급한 문제점들을 어느 정도 해결할 수 있다. 태그 데이터를 구조화하기 위해서는 다양하게 수평적으로 존재하는 태그를 유사한 것끼리 묶어주는 방법을 이용할 수 있다. 유사한 개념을 표현하는 태그끼리 군집화하여 이용자에게 제공한다면, 태그 공간의 다양성으로

인한 언어의 중의성 문제가 어느 정도 해소되어 탐색의 효율성 저하를 막을 수 있고, 유사한 개념으로의 탐색 확장도 보다 편리하게 이루어질 수 있어 이용자로 하여금 보다 효율적인 태그 공간의 탐색을 가능하게 할 것이다.

2.2 용어 클러스터링과 태그 클러스터링

유사한 태그를 군집화하는 방법으로 정보검색에서 활용되는 클러스터링 기법을 이용할 수 있다. 클러스터 분석(cluster analysis)이란 데이터 집합이 갖고 있는 구조를 발견하는 것으로, 다차원 공간상에서 유사한 객체 집단을 식별하는 다변량 통계 기법 중 하나이다(Tombros 2002). 클러스터 분석은 문헌, 용어 등을 대상으로 다차원 공간에 존재하는 정보 집단을 유사한 객체끼리 자동으로 분류하기 위한 방법으로 응용되어 왔다.

특히 용어 클러스터링은 서로 관련 있는 용어들을 일정한 기준에 따라 모아서 여러 개의 용어 클래스를 형성하는 것으로, 정보검색에서의 용어 불일치 문제를 해결하기 위해 시작되었다고도 할 수 있다. 여기서 용어 불일치 문제란 동음이의어나 이음동의어 등과 같이 정보검색 환경에서 동일한 개념에 대해 문헌의 저자와 탐색자의 표현이 일치하지 않는 것을 의미한다(한승희 2004). 용어 클러스터링을 통해 해결하고자 하는 용어 불일치 문제는 다수 이용자 간의 태그 표현의 불일치 또는 다양성이라는 형태로 태깅 시스템에서도 발견되며, 이러한 문제점의 해소는 현재의 태그 기반 정보검색 환경에서 해결해야 할 문제와 유사하다. 이러한 관점에서 볼 때, 클러스터 분석을 태그 공간에 적용한다면, 비구조화된 태그 데이터를 대상으로 유사한 태그끼리 군집화할 수 있으므로 이용자로 하여금 효율적인 태그 공간의 탐험이 가능하도록 유도할 수 있을 것이다.

용어 클러스터링은 용어간의 통계적 연관성 분석에 기초한다. 용어간의 연관성은 용어의 동시출현빈도를 이용하여 측정하는데, 두 개의 용어가 많은 수의 문헌에 함께 출현하였다면 이 두 용어는 서로 관련이 있다고 보고 같은 클래스에 포함시킨다. 같은 원리로, 용어 클러스터링을 태그의 군집화에 적용했을 때 임의의 두 태그가 여러 문서의 태깅에 함께 사용되었다면 두 태그는 연관성이 있다고 볼 수 있다. 용어간의 통계적 연관성을 측정하는 유사계수는 크게 거리계수와 유사계수로 나누어 볼 수 있는데, 일반적으로 텍스트 데이터의 통계적 연관성을 측정하는 데 있어 거리계수보다 유사계수가 더 적합한 것으로 나타났다(Strehl, Ghosh, Mooney 2000).

일반적으로 클러스터링 기법은 비계층적 기법과 계층적 기법으로 나뉜다. 비계층적 기법은 미리 정해진 k개의 센트로이드를 중심으로 센트로이드와 객체와의 거리를 최소화할 때까지 n개의 객체를 k개의 상호배타적인 클러스터로 나누는 방식으로, 시간과 비용을 최소화할 수 있고 계산복잡도가 낮다는 장점을 가지고 있다. 그러나 클러스터링 결과가 k개의 센트로이드 선택에 따라 많은 영향을 받는다는 단점이 있다. 비계층적 기법에 속하는 알고리즘으로는 k-means 기법, 싱글패스(single pass) 기법 등이 있다(Willet 1988).

계층적 기법은 유사도가 강한 객체를 포함하는 작은 클러스터를 상대적으로 유사도가 낮은

객체를 포함하고 있는 좀 더 큰 클러스터에 포함함으로써 트리 구조로 데이터를 분류하는 기법이다. 계층적 기법은 객체가 처리되는 순서에 영향을 받지 않기 때문에 비계층적 기법과는 달리 클러스터링 결과가 안정적인 반면 비계층적 클러스터링 기법에 비해 처리시간이 길고 계산복잡도가 높다(Tombros 2002). 계층적 기법에 속하는 것으로는 완전연결 기법(complete linkage method), 단일연결 기법(single linkage method), 집단평균 기법(group average linkage method), 워드 기법(Ward's method) 등이 있다(Voorhees 1985).

완전연결 기법은 두 클러스터를 한 클러스터로 묶는 과정에서 클러스터를 구성하는 모든 객체쌍 간의 유사도가 가장 작거나 거리가 가장 멀도록 하는 방식으로 두 클러스터 간의 유사도나 거리를 계산한다. 반면 단일연결 기법은 두 클러스터를 한 클러스터로 묶는 과정에서 클러스터를 구성하는 모든 객체쌍 간의 유사도가 가장 크거나 거리가 가장 가깝도록 하는 방식으로 두 클러스터 간 유사도나 거리를 계산하기 때문에, 클러스터 내의 특정 객체는 같은 클러스터에 속한 객체와 가장 유사하다. 집단평균 기법에서는 두 클러스터 간의 유사도를 모든 객체쌍 간의 유사도 값의 평균으로 산출하기 때문에 클러스터 내의 객체간의 최소 유사도나 최대 유사도를 고려하지 않는다는 특징이 있다. 또한 워드 기법은 클러스터를 구성하는 객체 간의 유클리드 거리의 제곱오차를 최소화하는 방식으로 클러스터를 통합한다. 특히 워드 기법은 다른 계층적 기법에 비해 클러스터의 크기를 작고 균일하게 분류해주는 경향이 있기 때문에 용어나 개념의 자동분류에 적합하다고 알려져 있다(Ding, Chowdhury, and Foo 2001).

2.3 선행 연구

태그 공간을 효율적으로 탐색하기 위한 태그 클러스터링에 관한 연구는 비교적 최근에 들어와 많은 연구가 이루어지고 있다. 태그 클러스터링에 관한 초기의 연구는 Begelman, Keller, Smadja의 연구(2006)로, 이들은 Delicious에서 RSS를 이용하여 태그를 수집하고 그래프 이론에 기반한 스펙트럴 클러스터링(spectral clustering) 알고리즘을 이용하여 연관 태그를 클러스터링하였다. 특히 이 연구에서는 특정 태그와 의미적으로 강하게 연관된 태그를 식별하기 위해 빈도 분포가 급격하게 낮아지는 지점을 동시출현빈도의 절단점으로 결정하는 방법을 사용하였다. Shepitsen et al.(2008)은 특정 이용자가 태그를 부여하는 패턴에 근거하여 어휘의 의미모호성을 해소할 수 있다고 가정하고, 이를 기반으로 태깅 데이터에 클러스터링 기법을 적용한 개인화된 추천 시스템을 설계하였는데, 계층적 기법과 k-means 기법을 비교 실험한 결과 계층적 기법이 우수한 성능을 보인 것으로 나타났다. 또한, Simpson(2008)은 웹 기반 사회적 북마킹 서비스인 Delicious와 인트라넷 기반 사내 북마킹 서비스를 대상으로 태그를 클러스터링하고 시각화하는 연구를 수행하였다. 클러스터링 결과 특정 클러스터에 너무 많은 태그가 포함되는 결과를 나타냈는데, 이를 해결하기 위한 방안으로 사전에 덜 중요한 태그를 제외하는 클러스터링 자질 축소를 제안하였다.

클러스터링 기법 이외의 방법으로 태그를 계층화하거나 군집화하는 연구들도 있다. Candan, Caro, and Sapiro(2008)는 태그 클라우드를 대상으로 이들 간의 맥락 관계를 압축표현하면서 계층화하기 위해 잠재의미색인(latent semantic indexing)에 기반한 방법을 제안하였다. 이 기법을 통해 태그 클라우드를 구성하는 태그들 간의 숨어있는 의미 관계를 표현할 수 있게 되어 보다 효율적으로 태그 공간을 탐색할 수 있음을 확인하였다. 또한 Schrammel, Leitner, and Tscheligi(2009)는 태그 클라우드가 표현되는 방법에 따라 태그 공간의 탐색 성능이 달라질 것이라는 가정 하에, 태그 클라우드를 표현하는 방법을 네 가지, 즉 알파벳순, 무작위순, 폭소노미 기반, 언어학 기반으로 표현한 후 이용자가 특정 태그 및 그 태그와의 연관 태그를 찾아내는 데 있어 어떠한 방법이 가장 효과적인가를 비교・분석하였다.

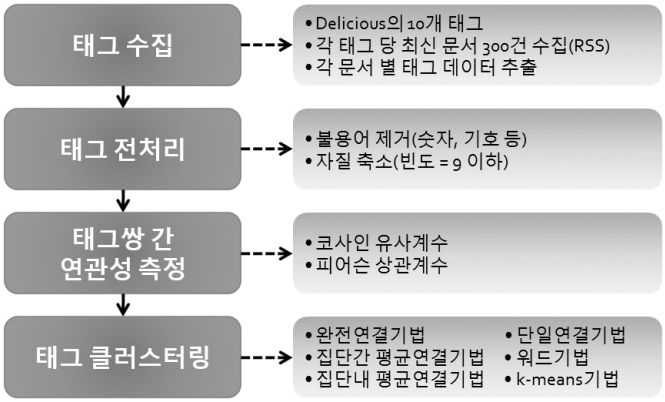
이외에도, 이순규, 김정훈, 이지형(2008)은 국내 블로그 서비스인 티스토리(Tistory, http://www.tistory.com)에서 수집한 5000개의 블로그 페이지를 대상으로 동시출현빈도 이외에 트랙백 페이지 간의 태그 동시출현빈도를 가중치로 적용하여 연관 태그를 클러스터링하는 방법을 제안하였으며, 이시화, 이만형, 황대훈(2008)은 Flickr에서 실험 데이터를 추출하여 효율적인 이미지 검색을 위해 태그 클러스터링을 활용하는 방안에 대해 연구하였다. 또한 Yi(2009)는 Delicious를 대상으로 연관 태그를 추출하는 방법에 대해 제안하였다.

3. 실험 설계

3.1 실험 개요

앞 절의 선행연구에서 언급된 Shepitsen et al.(2008)의 연구에서는 여러 계층적 기법 중 한 개의 기법(완전연결 기법)과 비계층적 기법의 성능을 비교한 후, 이 결과를 추천 시스템에 응용하는 연구를 수행하였다. 그러나 본 연구에서는 태그 공간의 효율적 탐색을 위해 연관 태그 클러스터의 형성에 어떠한 클러스터링 기법이 가장 효율적인가를 확인하기 위해 여러 클러스터링 기법을 비교하였으며, 이를 위해, 〈그림 2〉와 같이 (1) 태그 수집, (2) 연관 태그 추출을 위한 태그 전처리, (3) 태그쌍 간 연관성 측정, (4)태그 클러스터링의 네 단계의 과정을 거쳐 실험 결과를 획득하였다.

먼저, 본 실험을 위해 소셜 북마킹 서비스인 Delicious에서 임의의 태그 10개를 선정하여 2009년 7월 6일부터 8일까지 RSS를 통해 각 태그 당 최신 문서를 각각 300개씩 수집한 후 각 문서별로 태그 데이터를 수집하였다. 태그 데이터 별로 동일한 300개의 문서를 수집하는 데에는 상대적인 시간차가 있었다. 예를 들면, 상대적으로 다수의 이용자들이 관심을 가지고 있는 태그 'web 2.0'이나 'news' 등에 대해서는 300건의 문서를 수집하는 데에 하루가 소요되었으나, 상대적으로 문서의 등록 속도가 낮은 태그인 'food'나 'science' 등의 경우는 300건의 데이터가 수집되는 데에 3일이 소요되었다. 수집된 데이터의 통계적 특성은 〈표 1〉과 같다.



〈그림 2〉 실험 개요

〈표 1〉 실험 대상의 통계적 특성

항목 \ 태그	economics	education	food	library	music	news	photography	science	shopping	web2.0	평균
전체 태그 수	1288	1368	1096	1327	1208	1447	1219	1269	1196	1623	1304.1
문서 당 평균 태그 수	4.3	4.6	3.7	4.4	4.0	4.8	4.1	4.2	4.0	5.4	4.3
고유 태그 수	459	461	401	419	458	608	362	489	451	577	468.5

〈표 1〉에서 보는 바와 같이, 각 태그 당 300개의 문서에서 수집된 태그의 평균은 1304.1개이며, 문서 당 평균 태그 수는 4.3개로 나타났다. 이러한 데이터는, 사람들이 태깅할 때 한 문서 당 평균적으로 약 4개의 태그를 부여한다는 것으로 해석할 수 있다. 또한 전체 태그 중에서 중복해서 사용된 태그를 제외한 나머지 고유하게 출현한 태그의 수는 평균적으로 468.5개로, 실험집단 중 약 65%의 태그가 중복되어 나타난 것임을 알 수 있다.

수집된 태그 데이터 중 선정된 10개의 태그와 밀접하게 연관되어 있는 태그만을 추출하여 실험하기 위해 동시출현빈도가 10 이상인 태그만을 추출하였다. 동시출현빈도가 10 이상인 태그만 추출한 이유는 연관 태그를 수집하고 태그의 빈도분포에 대한 통계적 분석을 하는 과정에서 거의 모든 사례에서 빈도 10 근처에서 급격하게 빈도분포가 낮아지는 패턴을 보였기 때문이다. 그리고 이러한 패턴은 Begelman, Keller, Smadja(2006)의 연구에서 적용한 연관 태그 추출 방식과 유사하다고 할 수 있다.

각 태그 별로 최종적으로 추출된 연관 태그 수는 〈표 2〉와 같이, 평균 22.7개로, 가장 많은 연관 태그가 추출된 태그는 'web2.0'이고, 가장 적은 연관 태그가 추출된 태그는 'food'로 나타났다. 이를 대상으로 태그쌍 간의 동시출현정보에 기반하여 유사성을 측정하기 위해 태그×문서 행렬을 작성한 후 클러스터링을 수행하였다.

〈표 2〉 추출된 연관 태그 수와 생성된 클러스터 수

항목 \ 태그	economics	education	food	library	music	news	photography	science	shopping	web2.0	평균
연관태그 수	23	25	13	30	20	19	21	21	26	29	22.7
클러스터 수	6	6	4	7	6	6	6	7	7	7	6.2

본 실험에서는 정보검색 실험에서 일반적으로 널리 쓰이고 있는 코사인 유사계수(cosine coefficient)와 함께 동시인용 분석 및 동시출현 단어빈도 분석에 기초한 지식구조 분석 연구에서 많이 사용하고 있는 피어슨 상관계수(Pearson correlation coefficient)를 이용하였다. 태그 x와 태그 y에 대해 x_i는 문서 i에 출현한 용어 x의 가중치이며, y_i는 문서 i에 출현한 용어 y의 가중치일 때, 코사인 유사계수 $\cos(x,y)$와 피어슨 상관계수 $r(x,y)$의 공식은 다음과 같다(Sneath and Sokal 1973).

$$\cos(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

$$r(x,y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

$$(\overline{x} = \frac{1}{n}\sum_i x_i,\ \ \overline{y} = \frac{1}{n}\sum_i y_i,\ \ i = 1 \ldots n)$$

실험의 마지막 단계인 태그 클러스터링 기법의 단계에서는 계층적 알고리즘과 비계층적 알고리즘을 적용하였다. 계층적 알고리즘으로는 완전연결 기법, 단일연결 기법, 집단간 평균연결 기법, 집단내 평균연결 기법, 워드 기법을 적용하였고, 비계층적 알고리즘으로는 k-means 기법을 적용하였다.

각 실험집단 별 생성된 클러스터의 수는 휴리스틱을 이용하여 결정하였고, 그 결과 태그별로 4~7개의 연관 태그 클러스터를 생성하였다. 〈표 2〉에서 보는 바와 같이 생성된 클러스터 수의 평균은 6.2개이고, 한 클러스터 당 평균적으로 할당된 용어의 수는 약 3.7개였다. 연관 태그의 수가 클수록 생성된 클러스터의 수가 크고, 연관 태그의 수가 적을수록 생성된 클러스터의 수도 적어지는 것을 확인할 수 있다.

3.2 평가 방법

앞 절에서 언급한대로 다양한 조건의 클러스터를 생성한 후 어떠한 조건에서 연관 태그의 클러스터링이 가장 효과적으로 이루어지는지를 확인하기 위해 클러스터링 결과를 평가하였다. 클러스터링 결과의 평가는 기존의 혹은 사전에 작성된 분류정보와 클러스터링 결과와의 일치 정도를 측정함으로써 그 성능을 판정하는 것이 일반적이다. 그러나 본 연구에서 사용한 실험집단인 태그 데이터의 경우에는 사전에 작성된 분류정보가 없기 때문에 새로운 유형의 평가방법이 요구된다.

클러스터링 기법이 주제적으로 연관성 있는 태그들을 정확하게 군집화 해준다고 가정하면

한 클러스터에 속한 모든 태그들은 그 클러스터의 주제를 반영한다고 할 수 있다. 반면에 클러스터가 태그를 정확하게 군집화하지 못했다면 한 클러스터에 속한 용어들 중 일부는 그 클러스터의 주제를 반영하지 못한다.

그러므로 이 연구에서는 앞에서 언급한 가정을 근거로 이 실험에서는 주제적으로 연관성 있는 태그들을 군집화하는 클러스터링 기법의 효과를 측정하기 위해서 다양한 조건에서 생성된 클러스터링 결과를 대상으로 연구자가 직접 클러스터 적합도를 측정하였으며, 그 공식은 다음과 같다(한승희 2004).

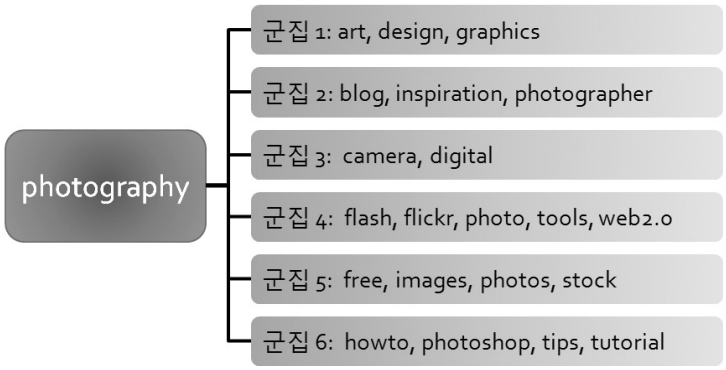
$$\text{클러스터 적합도} = \frac{\text{클러스터 대표주제에 적합한 용어 수}}{\text{클러스터에 속한 용어 수}}$$

클러스터에 포함된 태그들이 서로 주제적으로 연관성이 없는 경우에는 임의의 한 개 태그만 클러스터의 대표 주제에 적합하다고 간주하였다. 또한 한 클러스터에 두 개 이상의 주제가 함께 나타날 경우에는 더 많은 용어를 포함하는 주제를 클러스터의 대표 주제로 삼고 이를 기준으로 적합도를 평가하였다. 또한 한 개의 태그가 한 클러스터를 형성하고 있는 경우에는 적합하지 않은 것으로 평가하였다.

4. 실험 결과 분석

〈그림 3〉과 〈그림 4〉는 각각 태그 'photography'와 'library'에 대해 코사인 유사계수와 워드 기법을 이용하여 형성된 연관 태그 클러스터를 보여주고 있다. 그림에서 보는 바와 같이, 클러스터링 기법은 연관된 태그끼리 비교적 잘 군집화해주고 있는 것을 확인할 수 있다.

이용자의 태깅이 개인적인 정보관리의 목적에서 활용되는 것이 일반적이기 때문에, 실험결과에서도 특정 개념에 대해 유사한 합목적성을 갖는 태그끼리 군집을 이루는 방식으로 클러스터가 생성된 것을 확인할 수 있었다. 그렇기 때문에 연관 태그의 클러스터를 통해 연관 개념과 더불어 이용자의 정보이용행태와 선호관점 및 흥미분야 등을 확인할 수 있다. 예를 들면, 〈그림 3〉의 'photography'의 클러스터링 결과에서 군집 2는 사진작가의 블로그를, 군집



〈그림 3〉 태그 'photography'의 연관 태그 클러스터링 결과

3은 디지털 카메라와 관련된 정보를, 군집 6은 포토샵 활용을 위한 튜토리얼 관련 정보를 북마킹하기 위한 태그들이 한 군집으로 표현된 것을 확인할 수 있다. 이러한 클러스터링 결과가 이용자에게 태그 공간 탐색의 보조수단으로 제공된다면, 이를 통해 이용자는 같은 정보 공간에 있는 다른 이용자들의 선호관점이나 흥미분야를 참조하여 정보를 탐색할 수 있고, 자신의 정보 탐색의 목적에 맞게 원하는 태그를 추가하거나 개념을 확장하여 정보를 탐색할 수 있으며, 더 나아가 태그 검색의 장점이라 할 수 있는 우연한 발견을 더욱 활성화할 수 있게 될 것이다.

또한, 〈그림 4〉의 'library'의 경우에는 이 태그가 갖고 있는 중의성이 연관 태그를 통해서 해소되는 것을 확인할 수 있다. 주지하다시피, 'library'는 '도서관'과 '프로그래밍' 관련 개념을 가지고 있는데, 클러스터링 결과를 보면, 군집 1, 2, 3은 도서관과 관련된 태그들이 군집화되고, 군집 4, 5, 6, 7은 프로그래밍과 관련된 태그들이 군집화되어 있는 것을 확인할 수 있다. 이러한 결과를 통해, 태그 공간의 탐색에서 나타나는 언어의 중의성에 의한 문제점을 해결하는 데에 있어 태그 클러스터링이 어느 정도의 역할을 수행할 것으로 기대할 수 있다.

클러스터링 결과의 정확성을 확인하기 위해 다양한 조건에서의 클러스터 적합도를 평가한 결과는 〈표 3〉과 같다. 유사계수와 클러스터링 기법의 다양한 조합 중 코사인 유사계수와 워드 기법을 사용한 클러스터링 결과가 평균적으로 가장 좋은 성능을 보인 것으로 나타났다.

먼저 유사계수가 클러스터링 결과에 어떠한 영향을 미쳤는가를 살펴보면, 〈그림 5〉와 같이 워드 기법을 제외한 나머지 기법에서는 피어슨 상관계수가 코사인 유사계수에 비해 우수한 성능을 나타냈다. 그러나 유사계수의 차이가 클러스터링의 결과에 절대적인 영향을 미치지는 않는 것으로 나타났다. 워드 기법에서만 피어슨 상관계수가 좋지 않은 결과를 보인 것은 수학적 원리에서 찾아볼 수 있다. 워드 기법은 클러스터를 구성하는 객체간의 유클리드 거리의 제곱오차를 최소화하는 방법으로 클러스터를

〈그림 4〉 태그 'library'의 연관 태그 클러스터링 결과

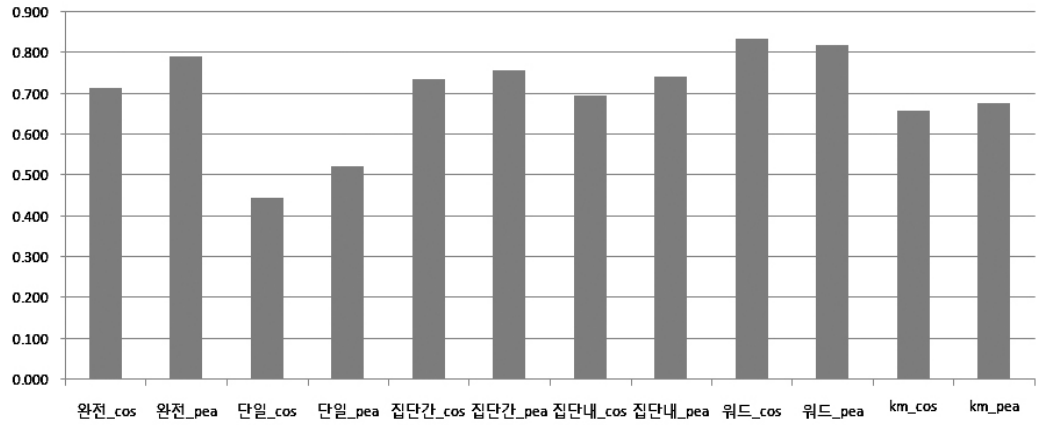
〈표 3〉 연관 태그의 클러스터링 적합도

기법 / 계수 / 태그	완전연결		단일연결		집단간 평균연결		집단내 평균연결		워 드		k-means	
	cos	pea	cos	pea	cos	pea	cos	pea	cos	pea	cos	pea
economics	0.522	0.522	0.217	0.217	0.478	0.565	0.522	0.522	0.870	0.652	0.522	0.565
education	0.680	0.680	0.440	0.560	0.800	0.840	0.760	0.840	0.840	0.840	0.600	0.600
food	0.769	0.923	0.538	0.538	0.846	0.846	0.923	0.923	0.769	0.846	0.692	0.769
library	0.800	0.933	0.567	0.667	0.833	0.900	0.800	0.867	0.933	0.933	0.700	0.767
music	0.800	0.850	0.400	0.700	0.750	0.750	0.750	0.750	0.850	0.850	0.650	0.700
news	0.474	0.632	0.316	0.158	0.526	0.526	0.526	0.579	0.737	0.737	0.579	0.632
photography	0.857	0.905	0.619	0.762	0.857	0.857	0.810	0.857	0.905	0.905	0.857	0.905
science	0.810	0.810	0.571	0.571	0.762	0.714	0.524	0.619	0.810	0.762	0.762	0.762
shopping	0.731	0.885	0.577	0.577	0.808	0.808	0.731	0.731	0.808	0.808	0.654	0.577
web2.0	0.690	0.793	0.207	0.483	0.690	0.759	0.621	0.724	0.828	0.862	0.586	0.483
평 균	0.713	0.793	0.445	0.523	0.735	0.757	0.697	0.741	0.835	0.819	0.660	0.676

생성하며, 이를 공식으로 나타내면 다음과 같다(Xu and Wunsch 2009).

$$E = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - m_k \|^2$$

워드 기법에서 취하고 있는 이러한 제곱오차의 최소화 원리는 피어슨 상관계수에서도 편차 제곱의 합의 형태로 나타난다. 즉, 편차 혹은 오차제곱의 최소화가 용어간 연관성을 측정할 때와 클러스터링을 수행할 때 각각 적용되기 때문에 오히려 이것이 성능을 향상시키는 데에 방해요소로 작용한 것으로 보인다. 그렇기 때문에 다른 기법에 비해 워드 기법에서만 유일하게 피어슨 상관계수가 좋지 않은 성능을 보인 것으로 해석할 수 있다.



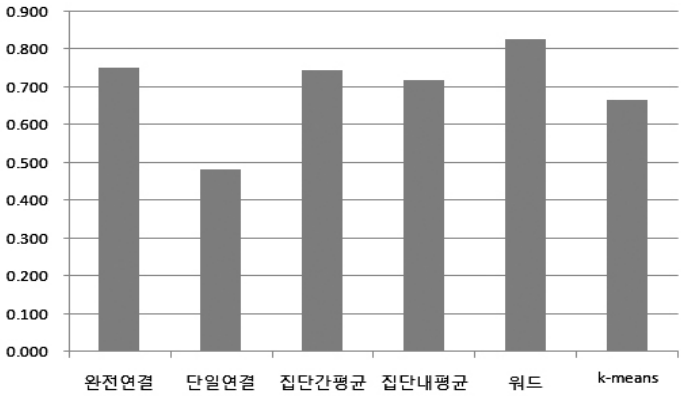
〈그림 5〉 유사계수별 클러스터 적합도 비교

클러스터링 기법의 측면에서 실험 결과를 살펴보면, 〈그림 6〉에서와 같이 단일연결 기법을 제외한 계층적 기법이 비계층적 기법인 k-means 기법에 비해 우수한 성능을 보인 것으로 나타났다. 또한 계층적 기법 중 가장 우수한 성능을 보인 기법은 워드 기법으로 확인되었다. 이 기법은 클러스터를 모두 비슷한 크기로 생성하는 경향이 있기 때문에 용어 클러스터링에 일반적으로 활용되고 있는데, 이 기법이 태그 클러스터링 결과에서도 가장 우수한 성능을 보였다는 것은 기존의 용어 클러스터링 방법을 태그 클러스터링에 적용하는 것이 가능하다는 것으로 해석될 수 있다.

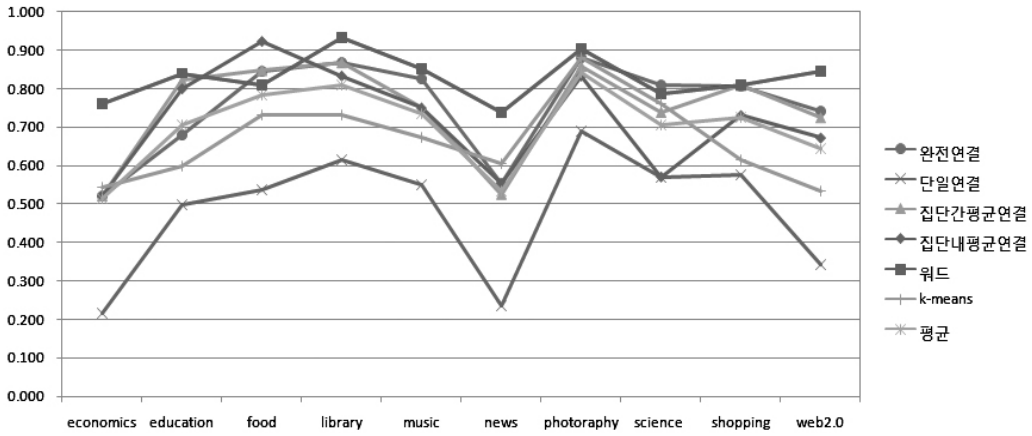
또한, 가장 성능이 좋지 않은 것으로 나타난 기법은 단일연결 기법으로 나타났다. 이 기법은 두 클러스터를 한 클러스터로 묶는 과정에서 클러스터를 구성하는 모든 객체쌍 간의 유사도가 가장 크거나 거리가 가장 가깝도록 하는 방식으로 클러스터를 생성하기 때문에 객체 간의 응집력이 약한 형태로 길게 늘어진 클러스터를 생성하는 경향이 있다고 알려져 있는데 (Jardine and Sibson 1968), 실험 결과, 특정 클러스터에 많은 태그가 집중되어 나타나거나 한 태그가 하나의 클러스터를 형성하는 등 태그 클러스터링에 적합하지 않은 것으로 확인되었다.

문헌 클러스터링에 적합한 것으로 알려진 완전연결 기법은 그 성능이 워드 기법 다음으로 우수한 것으로 나타났으나, 〈그림 7〉에서 보는 바와 같이, 실험집단 별로 성능에 편차가 있어 클러스터링 결과를 일관적으로 해석하기 어려운 것으로 나타났다.

비계층적 기법인 k-means 기법의 경우 계층적 기법 중 단일연결 기법을 제외한 나머지 기법보다 낮은 성능을 보였다. 단일연결 기법과 마찬가지로 일부의 실험집단에서 특정 클러스터에 많은 태그가 집중되거나 한 태그가 하나의 클러스터를 형성하는 경향을 보였다. 이러한 결과를 통해 k-means 기법이 용어 클러스터링에 적합한 기법이 아니라는 것을 확인할 수 있었다. 그러나 k-means 기법의 가장 큰 장점이라고 할 수 있는 간단한 계산복잡도는 대용량의 웹 데이터를 처리할 때 장점으로 작용할 수 있을 것으로 본다.



〈그림 6〉 클러스터링 기법별 클러스터 적합도 비교



〈그림 7〉 클러스터링 기법별 클러스터 적합도 추이

워드 기법과 집단간 평균연결 기법은 〈그림 7〉에서 보는 바와 같이 그 성능이 거의 실험집단에서 비슷한 경향을 보여, 이 두 기법이 태그 클러스터링에 일관적인 결과를 보여주는 것을 확인할 수 있었다.

5. 결론 및 제언

웹 2.0 환경에서의 태그 메커니즘은 정보의 조직에 있어 유동적이고 역동적인 분류체계의 지원을 통해 이용자 중심의 정보관리를 지원하고 있으나, 정보검색의 영역에서는 여러 한계점을 드러내고 있다. 본 연구에서는 태그 공간에서의 효율적인 정보 탐색을 위해 연관 태그를 이용자에게 효과적으로 제공하기 위한 방법 중 클러스터링 기법을 제안하고, 연관 태그의 클러스터 생성에 가장 적합한 기법을 확인하기 위해 다양한 클러스터링 기법을 적용한 후, 그 결과를 평가하여 비교·분석하고자 하였다.

실험 결과, 일반적으로 용어 클러스터링에서 성능이 우수한 것으로 알려진 워드 기법이 코사인 유사계수와 결합했을 때 모든 실험집단에서 유사한 경향을 나타내면서 태그 클러스터링에서도 가장 우수한 결과를 나타냈다. 클러스터링 기법별로는 단일연결 기법을 제외하고는 모든 계층적 기법이 비계층적 기법인 k-means 기법에 비해 좋은 성능을 나타냈다. 또한 유사계수별 특성으로는 워드기법에서만 코사인 유사계수가 나은 성능을 보였으며, 나머지 기법에서는 피어슨 상관계수가 우수한 성능을 보인 것으로 나타났다.

태그 클러스터는 일반적으로 이용자가 정보를 관리하기 위한 목적으로 부여한 태그를 군집화한 것이기 때문에, 그 결과 역시 합목적성이 있는 태그끼리 같은 군집에 속하는 경우가 많았다. 또한 일부의 클러스터링 결과에서 보여주는 바와 같이, 태그 클러스터링이 용어의 중의성을 어느 정도 해소할 수 있는 것으로 나타났다. 이러한 결과를 통해, 연관 태그 클러스터는 태그의 어휘적 관계를 반영하면서 동시에 이용자의 정보탐색 목적에 맞게 활용됨으로써

태그 공간의 효율적 탐색에 긍정적인 영향을 미칠 수 있을 것으로 예측된다.

태그 클러스터링에 적합한 클러스터링 알고리즘의 확인을 통해 태그를 활용한 검색 및 탐색, 연관 태그의 추천, 자동 태깅, 개인화 서비스 등 태그 공간의 효율적 탐색을 지원하는 방안을 개발하는 연구들이 더욱 활발하게 진행될 수 있을 것이다.

이 연구의 일반화를 위해 더 많은 실험집단을 대상으로 실험해 볼 필요가 있다. 일반적으로 태그쌍 간의 가중치는 바이너리의 형태로 나타나기 때문에 바이너리 형태의 가중치에 주로 사용되는 자카드 계수(Jaccard's coefficient)나 다이스 계수(Dice coefficient) 등과 같은 이진 유사계수를 적용하여 태그쌍 간의 연관성을 측정해볼 필요가 있다. 또한 태그쌍 간의 연관성 측정에 있어 용어 클러스터링에서 유효한 것으로 입증된 분포 유사도를 적용해 볼 필요가 있다(이재윤 2007). 분포 유사도는 두 객체의 확률 분포 사이의 차이를 측정하여 거리나 유사성을 판단하는 것으로, 이미 용어 클러스터링에서는 코사인 계수 대신 이를 이용하여 좋은 성과를 얻은 연구 결과가 발표된 바가 있다(Dagan and Lee 1999; Lee 1999; Weeds 2003). 또한 객체의 중복분류를 허용하는 퍼지 클러스터링 기법을 적용하여 클러스터링으로 보다 정교하게 용어의 중의성 문제를 해결하는 방안에 대해서도 연구가 필요하다.

참 고 문 헌

[1] 박병재, 우종우. 2008. 연관 태그의 군집 알고리즘의 설계 및 구현. 『한국IT서비스학회지』, 7(4): 199-208.

[2] 유사라. 1999. 『정보학연구와 분석방법론』. 서울: 나남출판.

[3] 이순규, 김정훈, 이지형. 2008. 트랙백을 이용한 연관태그 클러스터링. 『한국지능시스템학회 추계학술대회 학술발표논문집』, 18(2): 125-128.

[4] 이시화, 이만형, 황대훈. Web2.0 환경에서의 효율적인 이미지 검색을 위한 태그 클러스터링 시스템의 설계 및 구현. 『멀티미디어학회 논문지』, 11(8): 169-178.

[5] 이재윤. 2007. 분포 유사도를 이용한 문헌클러스터링의 성능향상에 대한 연구. 『정보관리학회지』, 24(4): 267-283.

[6] 이재윤, 정도헌. 2008. 폭소노미 태그 사용 패턴 분석 통제어휘 및 비통제어휘와의 비교. 『제15회 한국정보관리학회 학술대회 논문집』, 21-26.

[7] 이정미. 2007. 폭소노미의 개념적 접근과 웹 정보 서비스에의 적용. 『한국비블리아학회지』, 18(2): 141-159.

[8] 정영미. 2005. 『정보검색연구』. 서울: 구미무역(주)출판부.

[9] 정충영, 최이규. 2009.『SPSSWIN을 이용한 통계분석』. 제5판. 서울: 무역경영사.

[10] 한승희. 2004.『클러스터링 기법을 이용한 개별문서의 지식구조 자동 생성에 관한 연구』. 박사학위 논문, 연세대학교 대학원 문헌정보학과.

[11] Begelman, Grigory, Keller, Phillip, and Smadja, Frank. 2006. *Automated tag clustering: Improving search and exploration in the tag space*. [online]. [cited 2009.7.13]. 〈http://www.pui.ch/phred/automated_tag_clustering/〉.

[12] Candan, K. Selçuk, Caroz, Di, Luigi, and Sapino, Luisa, Maria. 2008. "Creating tag hierarchies for effective navigation in social media." *In Proceeding of the 2008 ACM Workshop on Search in Social Media*, 75-82.

[13] Dagan, Ido, Lee, Lillian, and Pereira, Fernando. 1999. "Similarity-based models of cooccurrence probabilities." Machine Learning, 34(1-3): 43-69.

[14] *Delicious*. [online]. 〈http://delicious.com〉.

[15] Ding, Y., Chowdhury, G. G., and Foo, S. 2001. "Bibliometric cartography of information retrieval research by using co-word analysis." *Information Processing and Management*, 37: 817-842.

[16] Fichter, Darlene 2006. "Intranet applications for tagging and folksonomies." *Online*, 30(3): 43-45.

[17] Hammond, Tony, Hannay, Timo, Lund, Ben, and Scott, Joanna. 2005. "Social bookmarking tools(I)." *D-Lib Magazine*, 11(4). [online]. [cited 2009.8.7]. 〈http://www.dlib.org/dlib/april05/hammond/04hammond.html〉.

[18] Jardine, N., and Sibson, R. 1968. "The construction of hierarchic and non-hierarchic classifications." *The Computer Journal*, 11(2): 177-184.

[19] Lee, Lillan. 1999. "Measures of distributional similarity." In Proceedings of 37th Annual Meeting of the Association for Computational Linguistics, 25-32.

[20] Mathes, Adam. 2004. *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. [online]. [cited 2008.7.31]. 〈http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html〉.

[21] Milligan, G. W., Soon, S. C., and Sokol, L. M. 1983. "The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure." *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 5(1): 40-47.

[22] Schrammel, Johann, Leitner, Michael, and Tscheligi, Manfred. 2009. "Semantically structured tag clouds: An empirical evaluation of clustered presentation approaches." *In Proceedings of*

the 27th international conference on Human factors in computing systems, 2037-2040.

[23] Shepitsen, Andriy, Janathan, Gemmell, Bamshad, Mobasher, and Robin, Burke. 2008. "Personalized recommendation in social tagging systems using hierarchical clustering." *In Proceedings of the 2008 ACM conference on Recommender systems*, 259-266.

[24] Simpson, Edwin. 2008. *Clustering Tags in Enterprise and Web Folksonomies*. [online]. [cited 2009.7.13]. 〈http://www.hpl.hp.com/techreports/2007/HPL-2007-190.pdf〉.

[25] Sneath, P. H. A., and Sokal, R. R. 1973. *Numerical Taxonomy*. SF: Freeman.

[26] Strehl, Alexander, Joydeep, Ghosh, and Raymond, Mooney. 2000. "Impact of similarity measures on web-page clustering." *In Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search(AAAI 2000)*, 58-64.

[27] Tombros, Anastasios. 2002. *The Effects of Query-based Hierarchical Clustering of Documents for Information Retrieval*. Ph.D. diss., Department of Computer Science, Cornell University.

[28] Voorhees, Ellen M. 1985. "The cluster hypothesis revisited." *In Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 188-196.

[29] Ward, Joe H. 1963. "Hierarchical grouping to optimize an objective function." *Journal of the American Statistical Association*, 58: 236-244.

[30] Weeds, J. E. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph. D. diss., University of Sussex.

[31] Willet, Peter. 1988. "Recent trends in hierarchic document clustering: a critical review." *Information Processing and Management*, 24(5): 577-597.

[32] Xu, Rui, and Wunsch II, Donald C. 2009. *Clustering*. NJ: IEEE Press.

[33] Yi, Kwan. 2009. "Mining semantically similar tags from delicious." *Journal of the Korean Society for Information Science*, 26(2): 127-147.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Byoung-Jae Park, & Chong-Woo Woo. 2008. "A Design and Development of A Related Tag Clustering Algorithm." *Journal of Korea Society of IT Services*, 7(4): 199-208.

[2] Sa-Rah Yoo. 1999. *Jeongbohakyeonguwa BunseokBangbeopron*. Seoul: Nanam Publishing House.

[3] Soon-Kyu Lee, Jung-Hoon Kim, & Jee-Hyong Lee. 2008. "Clustering Related Tag Using the

Trackback Mechanism." In *Proceedings of Korea Intelligence and Information Systems Society Fall Conference*, 18(2): 125-128.

[4] Si Hwa Lee, Man Hyoung Lee, & Dae Hoon Hwang. "Design and Implementation of Tag Clustering System for Efficient Image Retrieval in Web2.0 Environment." J*ounal of Korea Multimeda Society*, 11(8): 169-178.

[5] Jae Yun Lee. 2007. "Improving the Performance of Document Clustering with Distributional Similarities." *Journal of the Korean Society for Information Management*, 24(4): 267-283.

[6] Jae Yun Lee, & Do-heon Jeong. 2008. "Folksonomy Tag Sayong Pattern Bunseok Tongjee-ohwewaui Bigyo." *15nd Korean Society for Information Management HaksulDaehoi Nonmunjip*, 21-26.

[7] Jeong-Mee Lee. 2007. "A Conceptual Access to the Folksonomy and Its Application on the Web Information Services." *Journal of the Korean BIBLIA Society for Library and Information Science*, 18(2): 141-159.

[8] Young-Mee Chung. 2005. *Jeongbogeomsaekyeongu*. Seoul: Kumimuyeok.

[9] Choong-young Jung, & Ei Kyu Choi. 2009. *SPSSWINeul yiyonghan Tonggyebunseok*. 5nd ed. Seoul: Muyokgyeongyeongsa.

[10] Seung-Hee Han. 2004. *Automatic generation of the local level knowledge structure of a single document using clustering methods*. Ph.D. diss., Yonsei University.